DS Pallett

1995 HUB-3 NIST MULTIPLE MICROPHONE CORPUS BENCHMARK TESTS

David S. Pallett, Jonathan G. Fiscus, William M. Fisher, John S. Garofolo,
Alvin F. Martin and Mark A. Przybocki

National Institute of Standards and Technology (NIST)
Room A216, Building 225 (Technology)
Gaithersburg, MD 20899
Email: dpallett@nist.gov

ABSTRACT

This paper reports on large-vocabulary continuous "read" automatic speech recognition benchmark tests, and test materials, used in the 1995 Benchmark Tests using a NIST-collected Multiple Microphone Corpus. The tests used data collected with the by-now-traditional close-talking, noise-canceling Sennheiser HMD-410 microphone, as well as data simultaneously collected with three "secondary" microphones selected from a set of seven secondary microphones. Tests using the secondary microphones are termed P0 tests, and those using the close-talking microphone are termed the C0 ("contrastive") tests.

Most systems for which results are reported made use of some form of batch mode adaptation and/or channel compensation. Some systems, however, used strictly the same system for clean and noisy speech, others used arguably "different" systems. Differences included: different sets of acoustic models, different techniques for clean (speaker adaptation) and noisy speech (noise compensation), signal-property dependent differing weights for knowledge sources, differing usage of gender-dependent or gender independent models and different decoding strategies for clean and noisy speech. This variety of interpretations of [the term] "same system" results in a pretty wide and interesting set of results..." [1]

The test data were collected in an ambient noise environment measured at ~ 55 dB (A-weighted). NIST's measured SNR values for the secondary microphones included in these tests are typically in the range 10 to 20 dB, broadband, or 18 to 21 dB A-weighted. In contrast, SNRs of about 38 dB, broadband or A-weighted, are found for the close-talking microphone.

Word error rates reported for the secondary microphone data set (P0) range from 13.5% to 55.5%, and word error rates for the close-talking microphone data set (C0) range from 6.6% to 20.2%. For many systems, error rates for the secondary microphone data set are nearly double those for the (primary) close-talking microphone.

A number of factors affecting word error rate have been investigated at NIST and are discussed.

1. INTRODUCTION

It is well known that the performance of Automatic Speech Recognition (ASR) technology degrades when there is competing noise along with the speech signal. It is also well known that performance degrades when there are differences between the signal transmission channel(s) and/or microphones that are used for system acoustic-model-building and the channels and/or microphones used in tests or applications. Previous ARPA-sponsored benchmark tests included some studies of these phenomena, notably in the 1993 "Spoke 5 -- Microphone Independence", "Spoke 7 -- Noisy Environments" and "Spoke 8 -- Calibrated Noise Sources" tests [2], and in the 1994 "Spoke 10" tests involving the use of additive automobile-interior noise [3]. The 1993 studies involved the use of only two microphones at any one time ("stereo" data) -- thus limiting the amount of data from any one microphone in any one noise environment. This factor also affected the statistical significance of findings obtained with these data. The 1994 tests were criticized because the use of additive noise (added to otherwise "clean" speech) in general, may not simulate the properties of speech data collected in a "real" noise environment.

Some initial commercial implementations of ASR have resulted in less-than-satisfactory performance and limited user acceptance of ASR technology because of performance degradations that occur with the use of inexpensive, often sub-optimally positioned, microphones.

With funds provided by the Department of Commerce's Advanced Technology Program in Fiscal Year 1994, NIST developed an eight-channel A/D speech data collection system to enable the study of channel- and microphone-related effects by researchers and technology developers. We refer to this MUltiple Microphone speech data collection system as the NIST "MUM data collection system".

Early in 1995, a small pilot corpus consisting of "read" Wall-Street-Journal speech was collected at NIST in a moderately "live" (reverberant) computer laboratory with an A-weighted sound level of ~ 55 dB and a broad-band (unweighted) sound level of ~ 70 dB. That A-weighted level falls in a range that Beranek cites for "moderately fair listening conditions" with "steady background noise" ranging from 52 to 61 dBA for "light maintenance shops, office and computer equipment rooms, kitchens and laundries. Beranek also indicates that this level represents noise levels ~ 14 dB higher than might be typical for private and semiprivate offices, small conference rooms, classrooms, and libraries (A-weighted levels in the range 38 -47 dBA). [4]

This pilot corpus was used to investigate the performance of an HMM recognizer trained with speech data collected with the use of conventional close-talking noise canceling microphones (e.g., Sennheiser HMD-410, or equivalent). Eight microphones, in all, were used in the NIST investigation, including the conventional close-talking microphone, the output of a high-quality precision sound level meter placed slightly more than a meter from the speaker, three high quality microphones, and three inexpensive microphones. In analyzing the results of the ASR experiments conducted with these data, it became clear that while the data collected with the close-talking microphone yielded near-state-of-the-art performance; substantially higher error rates were observed using the data from the other microphones. These results were as we expected, and are attributable to differences in microphone and/or channel properties and the higher levels of interfering noise in the data from the seven "secondary microphones".

It appeared obvious that the use of any of several forms of channel-compensation and adaptive techniques would reduce error rates, but it was unknown how significant the reductions might be using state-of-the art techniques. Government sponsors expressed interest in having NIST make benchmark test materials available for this purpose to researchers in this community, and proposed that a working group be formed to outline an appropriate test paradigm. This paper presents NIST's summary and tabulations of the results submitted to NIST in November, 1995, as well as some commentary on the properties of the data and results.

2. HUB-3 TEST MATERIALS

The data used for these tests was collected at NIST, with funding provided by the Department of Commerce. All data were collected with the NIST MUM data collection system described in Appendix 1. These materials will be made available to the general public in the near future through the Linguistic Data Consortium.

2.1 NIST MUM Development Test Data

A development test set consisting of recorded speech from 20 subjects reading prompting texts that were identical to last year's Hub-1 development test set, was collected at NIST. The NIST MUM data collection system was positioned at two locations within a NIST laboratory for this development test set. The development test set was provided by NIST to potential test participants in mid-summer 1995.

2.2 1995 CSR Hub-3 Evaluation Test Data

Like the development test set, the evaluation test set consisted of 20 subjects. The subject population for these data was drawn from NIST technical and administrative staff, all of whom are native speakers of American English. The data are balanced for sex (10 males/10 females). Each subject first spoke the same 6 NAB "warm-up" sentences (Session 1), and then (nominally) 15 unique NAB sentences (Session 2) drawn from 1 of 20 different news articles selected by NIST from several August 1995 news sources. The two sessions were recorded back-to-back, without an appreciable break between the two sessions. A 5 second "background noise" recording was also made at the beginning and end of each of the 2 sessions.

The test material that was to be processed in these benchmark tests (which did not include the 6 NAB "warm-up" sentences) comprised, in all, two sets of nominally 300 utterance files -- 20 speakers, each speaking 15 utterances. One set was collected using the "primary" close-talking, noise-canceling Sennheiser HMD-410 microphone, and was used for the C0 tests. The other set consists of three subsets of material collected simultaneously, each subset consisting of material collected from one of three selected microphones. These comprise a subset of the 2400 utterance file set collected with the MUM data collection equipment (20 speakers X 15 utterances/speaker X 8 microphone files/utterance).

One secondary microphone "subset", Set 1, with 7 speakers, was obtained using a Shure boom-mounted SM58 cardioid dynamic microphone. Set 2, with another 7 speakers, was obtained using an Audio Technica AT851a "micro cardioid condenser boundary" microphone, and Set 3, with another 6 speakers, was collected using a relatively inexpensive Radio Shack #333-1060 "omni electret" microphone.

Each of the 7-speaker subsets consisted of 105 utterance files. The 6-speaker subset consisted of 90 utterance files. Note that these individual test subsets are small, which is a factor that may limit the statistical validity of some of the conclusions of these experiments. Larger secondary microphone test sets, of course, could have been defined -- potentially including all data for all 300 utterances from each of the 7 secondary microphones (a 2100 utterance file secondary test set). Alternatively, the entire 300 utterance file sets for each of the 3 selected secondary microphones could have been used. But participants in the tests argued that secondary test sets this large would require excessive computational time. NIST has used all of the data in unpublished studies.

2.3 Data Collection Environment

The data were collected in a somewhat reverberant 11 X 24 foot laboratory/office module. This module has floor-to-ceiling metal partition walls, carpeted raised computer-room floors, an 11-foot high textured concrete ceiling, and metal doors on three sides. It contained several workstations, disk drives, and printers and a CD-ROM "jukebox". The ambient noise near the data collection workstation (and microphones) in the room was measured at 72 dB (linear) and 54 dB (A-weighted). The same room was used for collection of both the development test and evaluation test materials.

All of the evaluation test data were collected at one location within the room. The location was at approximately the mid-point of one of the long walls, as was the case for one of the locations used in collecting the development test data, but the data collection system setup differed in that a different wall was used and there were no acoustically absorptive surfaces or panels in the immediate proximity.

2.4 Test Set Text Selection Procedure

As stated in the specifications for Hub-3, approximately equal numbers of articles were selected from each of five sources (Wall Street Journal/Dow Jones Industrials Service, New York Times, Reuters/Reuters North American Business, Los Angeles Times, and the Washington Post.

Fairly large samples of articles from each of the sources for the month of August, 1995 were processed through LDC tools that produced tagged and sentence-labeled versions. One step in this process that was quite time-consuming was manual correction of sentence-boundary determination. For the sources that had more raw data than could be handled, this was done on a subset of August files selected more-or-less at random, but fairly evenly over the month. All articles with fewer than 15 sentences were then deleted from this pool. An attempt was made to find and delete duplicate or near-duplicate articles. The pool of articles was further winnowed by filtering out articles that had a very long or a very short sentence among the first 15.

Of the pool of remaining articles -- on the order of 40 for each source -- 15 articles were picked at random from each source and a prompt set made for each, containing only its first 15 sentences. Each of these 15 articles, per source, was then examined and a subjective judgement was made as to whether it would be too hard for a typical subject to read. If it was judged too hard, another article from the remainder of the pool of about 40 per source was picked. Some reasons for rejecting articles were: (1) obvious garbling of the texts, (2) graphic devices that we didn't know how to pronounce, (3) words that might embarrass the reader, (4) tabular data, (5) meta-text, and (6) an extremely high density of foreign names. Other than this culling, no attention was paid to balancing the perplexity of sentences in the different prompting text sets.

Of the 15 articles ultimately selected from each of the five sources, four were assigned at random to each of three 20-article test sets, leaving the extra three to be held in reserve in case we discovered some hidden flaw in an article selected for a test set. Each of the 20 unique 15-sentence articles was then read by one of the 20 subjects, preceded by a reading of a short 6-sentence article for practice, which was the same 6-sentence article for all speakers.

One of the three 20-article test sets was arbitrarily selected for use in the 1995 Hub-3 test set, leaving two others for future use.

3. HUB-3 TEST PARADIGM

The Working Group agreed that the goals of Hub 4 were to:

> improve basic SI performance on unlimited-vocabulary read speech under acoustical conditions that are somewhat more varied and degraded than speech used in previous evaluations. The evaluation data set [was to] include data recorded using several different microphones simultaneously. One of these microphones [was to] be the standard "close talking" microphone that had been used in previous ARPA spoken language evaluations. In addition, several "other", generally non-close talking, microphones [were to] be included in the evaluation data. These "other" microphones [were] not [to] have been used in training sets, development test sets, or evaluation test sets in previous ARPA evaluations. All speech [was to] be recorded in a room with background noise in the range of 47 to 61 dB (A weighted).

Please see another paper in this Proceedings for additional information about the test specifications [5].

Participants in these tests included eight groups at seven sites: AT&T Bell Laboratories [6], CMU [7], Cambridge University Engineering Department (both Connectionist [8] - and HMM [9] - systems), LIMSI [10], New York University [11] (in collaboration with SRI International), Rutgers University [12], and SRI International [13]. All of these sites, except for Rutgers, have participated in previous ARPA-NIST CSR benchmark tests.

As in previous tests, sites were permitted to submit "requests for adjudication" after the preliminary scoring. During adjudication, sites submitted bug reports to NIST via Email. These adjudication bug reports usually request that a transcription be "corrected" or that an alternate transcription be permitted. Sites were also permitted to comment on, or contest, other site's requests. This process, although laborious, ensured that the transcription of the test data and resultant scored results were as accurate as possible.

This year, for the Hub-3 tests, NIST received 316 requests pertaining to 110 unique utterances. The 316 requests were serviced as follows: 156 - denied, 135 - granted, 21 - partially granted, and 4 - no action (general comment only, no request actually made).

Note that there were many requests for changes to compound or hyphenated words which were already handled by a general hyphenation rule or a global mapping file used to forgive certain inconsistencies in the hypothesis and reference transcriptions. So many requests which were "granted" resulted in no change. There were actually very few "real" transcription errors - on the order of 10 words out of the 6025 words in the test set. Most changes to the transcriptions involved the addition of an alternation due to acoustic and contextual ambiguity. Even these were relatively few. In all, changes were made to only 22 of the 300 single-utterance transcriptions. The following is the breakdown of the changes to the transcriptions: word correction - 3, word deletion - 2, word addition - 2, word fragment addition - 1, compound word changes affecting the lexical SNOR transcriptions - 2, alternations - 10. An additional 11 word mappings (some with multiple forms) were added to the global map file to handle ambiguous compounding cases, spelling variants, and cases where the language model disagreed with the rules. These additional rules affected only 12 of the 300 single-utterance transcriptions.

It should be noted that the Hub-3 adjudication resulted in a decrease in error rate of only about 0.2% in most systems' results and resulted in no change to the relative ranking of the different systems.

4. HUB-3 TEST RESULTS

Table 1 presents NIST's tabulation of the Hub-3 results submitted to, and scored by, NIST. Results are tabulated for twelve systems from eight sites. These include two systems (att1 and att2) developed at AT&T (including one that was submitted late in a "debugged" version), two from CMU (cmu2 and cmu2b), a "connectionist" system developed at Cambridge University's Engineering Department (cu-con1), an HMM system also developed at Cambridge University (cu-htk1), three systems developed at LIMSI (including one that was submitted late in a "debugged" version), and one system that represented a collaborative endeavor involving NYU and SRI International (nyu-sri1), one from Rutgers University (ru1), and one from SRI International (sri1). For some sites, results were submitted for "contrastive" conditions; others submitted only one set of results.

All sites were required to process data from two 300 utterance-file sets. One set of data, derived from the primary, close-talking, noise-canceling Sennheiser HMD-410 microphone, was used for the C0 "contrast" condition, intended to contrast with the results for the secondary microphones. The aggregate set of secondary microphone data from 3 microphones was used for the P0 test conditions. (This was referred to, perhaps unfortunately, as the "primary" test condition (P0) because it emphasized the use of alternative microphones, despite the fact that these microphones are described as "secondary" microphones.) The test specification indicated that the "same system [was to be] used for both P0 and C0". The P0 and C0 tests were intended as "open baselines" and could employ different forms of compensation, and/or adaptation although the "same system" was to have been employed for both tests. Data were provided for only a few contrastive condition, for comparison with these open baselines.

Most systems for which results are reported made use of some form of batch mode adaptation and/or channel compensation. Some systems, however, used strictly the "same system" for clean and noisy speech, others used arguably "different" systems (e.g., different sets of acoustic models, different techniques for clean (speaker adaptation) and noisy speech (noise compensation), signal-property dependent differing weights for knowledge sources, differing usage of gender-dependent or gender independent models and different decoding strategies for clean and noisy speech. This variety of interpretations of [the term] "same system" results in a pretty wide and interesting set of results..." [1]

One site (Rutgers) did not implement any form of adaptation to the test data. In Table 1, their data are shown in both the P0 and C0 columns although it more appropriately belongs (only) in the relevant columns for constrastive conditions (C1A and C1B) that were intended for static "controlled evaluation tests" using standardized training sets and 60K word trigram grammars. Another site (AT&T) invoked only minimal language model adaptation for the P0 and C0 tests.

Note that for the P0 condition, involving the secondary microphones, the word error rate ranges from a low of 13.5% (for the cu-htk1 system) to 55.5%. In contrast, for the C0 condition, for the data from the close-talking microphone, the word error rate ranges from a low of 6.6% (also for the cu-htk1 system) to 20.2%.

Table 1 also presents summary tabulations of significance tests implemented by NIST on within-site results. Note, for example, that for the cu-htk1 system, when comparing the error rates for the two conditions, the error rates for the secondary microphone set are increased by 103.5% relative to those for the close-talking microphone. Other within-site comparisons typically yield increases in the word error rate for the P0 and C0 data ranging from 100% to 150% -- that is, the error rates for the secondary microphone set of data are typically 2 to 2 ½-times larger than for the close-talking microphone.

Tables 2 and 3 present the results of NIST's (now traditional) implementation of several paired-comparison statistical significance tests. By convention, when the null hypothesis is not rejected, we print the word "same" to suggest that the word error rates (or the utterance error rates, in the case of the McNemar "MN" test) are not (shown to be) significantly different. By the same convention, in the case that the null hypothesis is rejected, we print the identity of the system with the lower error rate.

Cross-site comparisons for these tests, such as those documented in Tables 2 and 3, may not be particularly scientifically informative because of differing interpretations of the test specifications.

Table 4 presents NIST's tabulation of the Hub-3 results submitted to, and scored by, NIST, with a more detailed breakdown of results into the three subsets of data from the three microphones. Note that of the three data (sub-) sets, the highest error rates are reported for "set3", which was obtained with the inexpensive electret microphone. Note also that in general, and for both sets of data, higher error rates are reported for "set3" than for "set1" and "set2". In particular, for the C0 tests, this is believed to be attributable to the small size of the data subsets and the differences in perplexity for the texts chosen for these subsets, as discussed later.

In consideration of the small size of these data subsets, and the observed non-uniformity of properties of the data subsets, we have chosen not to provide tabulations of the results of paired-comparison significance tests.

5. HUB-3 DISCUSSION

There are, of course, a number of measurable properties of the test data. For each test set or subset these include: (1) the perplexity of the read texts (computed using a reference language model); (2) the rate of incidence of Out-Of-Vocabulary (OOV) word tokens ("new words"), expressed as a percentage of all word tokens in the test (sub-)set; (3) the number of word types (as opposed to tokens); (4) any of several measures of rate of speech; and (5) measures of the signal-to-noise-ratio (SNR) of the test material.

Table 5 documents some of the measured properties of the Hub-3 test material. Data are presented for each of the 20 speakers as well as appropriate mean values for Subsets 1, 2 and 3 and the overall test set.

Note that for the test subsets spoken by individual speakers, the measured mean sentence perplexity for the test subset (using the 60 K trigram language model developed by Rosenfeld [14]) varies widely, from a low of 10 for subject 71c, to 964 for subject 71f. The texts of these two subjects' articles are included as Appendix 3.

The rate of OOVs for all speakers in all subsets also varies appreciably -- from 0.0% to 4.1%.

"Speaking rate" has previously been identified as an important factor affecting the performance of ASR technology, and several different measures of this property have been used. In prior years [2, 3, 15], NIST used simple measures of speaking rate by counting the number of word tokens in each speech file, and dividing this by the total duration of the files. Rates typically reported range from 120 to 200 words/minute (~ 2 to ~ 3 words/sec) . This measure is affected by the (variable duration) time before speech is initiated within the individual "utterance" files, as well as that after the cessation of speech. It is also affected by the duration of pauses within the files. Another approach involves segmentation of the files, and accounting for only the "speech" time. A third approach makes use of information that can be derived from the use of segment time-marked files, where information on segmentation can be at the syllable or even at the phone level.

Another set of quantitative analyses of the variables that might be affecting word error rate is presented in Fisher's paper in this volume [16] Note that speech rate, in these studies, is typically in the range of 180 - 210 words/minute.

NIST's routine analyses of speech and noise levels make use of measures of peak speech power levels and mean noise power levels, using an approach that is based on analyses of power histograms and uses the 95'th percentile level for the residual power, after removal of data attributable to the noise. "Peak speech power levels" are, of course, higher than the RMS speech power levels. In a series of experiments with the data of Hub-3, NIST found that, in most cases, the measured "peak speech power levels" exceed the RMS speech power by approximately 5 to 6 dB (the range is 4.5 dB to 6.0 dB, with subset standard deviations typically in the range of 0.6 dB to 1.3 dB). So, if one wishes to make comparisons with other reported SNR measurements which make use of both mean noise power levels and mean speech power levels, one must subtract 5 or 6 dB from NIST's reported SNR data.

In many studies of noise phenomena, use is made of frequency-weighting schemes such as the standard A-weighting characteristic, which de-emphasizes information at low frequencies in a manner that attempts to simulate the "inverse 40 Phon" contour derived from psychophysical considerations. For signals with significant low-frequency noise, higher values are typically reported for A-weighted SNRs, typically 5 to 6 dB higher than for the unweighted values, depending on the spectral distributions of the speech and noise.

Both unweighted and A-weighted SNR data are tabulated for the test set data in Table 5. Note that values of ~ 38 dB (unweighted or A-weighted) are typical for the close talking microphone's data. Note also that for the data in the three subsets corresponding to the three secondary microphones (in the three test subsets), the unweighted SNRs are 18.9, 20.5, and 10.1 dB, respectively, while the corresponding A-weighted SNRs are 20.7, 21.2, and 18.2 dB. The difference in unweighted and A-weighted SNRs for the "g" microphone suggests the presence of substantial low-frequency noise for this microphone.

NIST's analyses of the data for Hub-3 have included both simple, qualitative, and more complex quantitative analyses to investigate factors for which there appears to be strong correlation with increased error rate, in addition to speaking rate, however measured. Fisher's paper [16] presents preliminary results of NIST's quantitative multi-dimensional analysis of variance.

Figure 1 illustrates a qualitative approach that suggests that there is strong correlation of word error rate with some of these factors. The data reported on in Figure 1 were derived from the results reported for the Rutgers system, which does not incorporate adaptation. The data in the lower half of the figure were derived from the close talking Sennheiser HMD-410 microphone, and the material in the upper half was derived using data from the three secondary microphones -- the Shure SM58 microphone for Test Subset 1, the Audio Technica AT-851a microphone for Test Subset 2, and the Radio Shack 33-1060 electret microphone for Test Subset 3.

The data shown in the lower portion of Figure 1, for Subset 1, indicate the word error rates for the 7 different speakers in this subset of the (C0) test data, using the close-talking microphone, ranging from ~4% to ~22%. In small windows above these data, corresponding data are plotted for the individual speakers': test subset's mean utterance perplexity, %OOV, speaking rate in words per minute (using forced alignment), and measured unweighted signal-to-noise ratio (here plotted with a scale with 10 dB at the top and 50 dB at the bottom). Note that there appears to be evidence of correlation between error rate and perplexity and the percentage of OOVs. In contrast, speech rate and SNR do not appear to be so obviously correlated with error rate for the C0 test data..

The upper portion of Figure 1 presents a similar display for the data derived from the same utterance and speaker set, but using the (P0) Shure SM58 secondary microphone. Here there is also a generally increasing error rate, ranging in this case, however, from ~18% for subject 716 (on the left) to ~50% for subject 71j (on the right). The maximum error rate is ~70% for subject 71c. For this subject, however, note that the SNR appears to be less than for other speakers in this set. Note also, that it may be the case that rate of speech is somewhat lower than for other speakers in this subset. The evidence suggests that the reduced SNR may be responsible for degraded performance, at least for speaker 71c, but it is far from persuasive.

Next, consider the data for the other two subsets of speakers in Figure 1(c) and 1(d) (for subset 2), and Figure 1(e) and 1(f) (for subset 3). In figures 1(c) and 1(e), the (C0) data for the Sennheiser microphone were used. For figure 1(d) the (P0) data for the Audio Technica AT851a microphone were used, and for figure 1(f), the (P0) data from the Radio Shack electret microphone were used. Again we find qualitative evidence, as we might suspect from other studies, to suggest that perplexity may be the most important correlate of error rate for the data obtained with the close talking microphone, and that rate of speech may also be significant. For the two sets of data obtained with the secondary microphones, error rates are markedly higher. Referring to the data of table 2, one finds that the P0:C0 comparisons suggest increased error rates for the secondary microphones, relative to the primary microphone, in the three subsets of 166%, 171.9% and 176.9%. Unweighted SNR's for the primary microphones are ~ 40 dB for the primary microphone, and between 10 and 20 dB for the secondary microphones.

These displays serve to show evidence suggesting correlation of these factors with error rate, certainly for a system that does not perform adaptation.

Figure 1 presents a similar display of results for the well-performing cu-htk1 system. Note again that there is evidence of correlation of error rate and test set perplexity (and OOV rate) even for the close talking microphone (C0) data. However, note that error rates for the secondary microphone (in the P0 test) are still significantly higher than for the primary microphone (in the C0 test).

Because of the small size of these test sets -- a total of 20 articles and 20 speakers -- and even smaller test subsets for the individual secondary microphone subsets -- 6 or 7 speakers, comprising 90 or 105 sentence utterances -- and limited variations in some parameters such as SNR within any one subset, it may be the case that only gross trends can be identified. However, further implementation of ANOVA techniques may suggest better experimental designs for future tests.

6. SUMMARY AND CONCLUSIONS

The NIST Hub-3 MUM corpus was designed to make possible direct comparisons of error rates for simultaneously recorded material from a number of secondary microphones with error rates obtained using a close-talking microphone. The test data were collected in a 55 dB A-weighted sound level environment.

Most systems in these tests used a number of adaptive techniques and procedures for channel compensation, sometimes based on the use of alternative models depending on measurements of the apparent attributes of the channel or noise, and in some cases using multiple passes. At least two sites did not implement adaptation, or did so minimally.

For systems that did not use adaptive techniques, word error rates for the secondary microphone test set were several times higher than for the primary, close-talking, microphone. For the best-performing of the adaptive systems (based on lowest overall word error rate on both test sets), the error rates for the secondary microphone set were approximately double that of the close talking microphone (i.e., comparing both the P0 and C0 test results indicates that there is an increased word error rate of 103.5% for the P0 results, relative to the C0 results).

Lowest word error rates for the close talking microphone data set, 6.6%, are slightly less than last year's lowest word error rate, on last year's (Hub-1 P0) test set, despite procedures for selection of test set texts from more diverse sources and higher ambient noise in the data collection environment.

The ambient noise environment in which the Hub-3 data were collected, while higher than in previous tests, is not severe by any measure -- Beranek describes this range as for "moderately fair listening conditions". The SNR of the close talking microphone's data is ~ 38 dB. NIST's measurements of the subset mean values of A-weighted SNR values for the three secondary microphones in this test set are in the range of 18.2 to 21.2 dB, using NIST's measures of "peak speech power" levels. If RMS measures of speech level were used to estimate SNR, these values would be approximately 6 dB less. The ambient noise environment, principally due to mechanical and HVAC sources, is also remarkably stationary.

NIST has conducted analyses involving measurable factors that may correlate with error rate. This study has included consideration of the average perplexity of the read test set material, the rate-of-incidence of Out-Of-Vocabulary (OOV) word tokens, speech rate, and SNR properties of the test material, as well as microphone type. In a related study at NIST, Fisher has shown that both fast and slow speech may affect word error rates [16].

These studies suggest that for current technology, given only a limited amount of material for adaptation (in this case, at most 15 sentence utterances), word error rates from secondary microphones can be expected to be double (or possibly quadruple) those found for close-talking, noise canceling microphones.

NIST plans to collect and release more MUM speech data in different noise and acoustic environments, and intends to make it possible to repeat this Hub test with new test data next year.

7. ACKNOWLEDGMENTS

Funds for the development of the MUM data collection system were provided by the Department of Commerce's Advanced Technology program. At NIST, Sean Sell was responsible for development of the MUM data collection system. Several individuals (including Jon Fiscus, John Garofolo, Mark Przybocki and Mike Walker) participated in collecting data from more than 40 cooperative volunteer speakers. Texts used for the recorded utterances were made available by the Linguistic Data Consortium, under several licensing agreements. The Hub-3 experimental paradigm was developed by Richard Stern at CMU, working with representatives of the participating sites and communicating with research sponsors and ARPA.

NOTICE

THIS PAPER IS A DRAFT VERSION AND AS SUCH IS SUBJECT TO CHANGE PRIOR TO THE FINAL PUBLICATION IN THE PROCEEDINGS OF THIS WORKSHOP.

Certain commercial equipment, instruments, or materials are identified in this document in order to adequately specify experimental procedures used. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] Gauvain, J. L., personal communication (email), Dec. 6, 1995.

[2] Pallett, D. S., et al., "1993 Benchmark Tests for the ARPA Spoken Language Program", in Proc. Human Language Technology Workshop, March 8-11, 1994, Plainsboro, NJ.

[3] Pallett, D. S., et al., "1994 Benchmark Tests for the ARPA Spoken Language Program", in Proc. Spoken Language Systems Technology Workshop", Jan. 22-25, 1995, Austin, TX.

[4] Beranek, Leo L., "Noise and Vibration Control" New York: McGraw-Hill Book Company; 1971, pp. 584-586.

[5] Stern, R. M., (Hub-3 paper in this proceedings)

[6] (AT&T paper in this proceedings)

[7] (CMU paper in this proceedings)

[8] (Cambridge University, Connectionist paper in this proceedings)

[9] (Cambridge University, HTK paper in this proceedings)

[10] (LIMSI paper in this proceedings)

[11] (NYU/SRI collaborative effort, paper in this proceedings)

[12] (Rutgers University paper in this proceedings)

[13] (SRI International paper in this proceedings)

[14] Rosenfeld, R., (ref. to 60K trigram language manual)

[15] Pallett, D. S., Fiscus, J. G., Garofolo, J. S., "Resource Management Corpus: September 1992 Test Set Benchmark Test Results", in Proc. Continuous Speech Recognition Workshop, September 21-22, 1992, Stanford, CA.

[16] Fisher, W. M., "Factors Affecting Recognition Error Rate" (in this Proceedings).

APPENDIX 1.

NIST Multi-Microphone (MUM) Data Collection System and Procedures

The data were collected using a PC-based 8-channel direct-to-disk recording system and NIST-developed data collection software.

Eight different channels (microphones) can be used as input to the system, via a mixer. A professional quality Mackie 16 - 8 Bus Console Mixer was used to interface the 8 microphones to the A/D equipment and control recording gains on all channels. All 8 microphones were connected to the mixer. The gains were independently adjusted to provide comparable speech audio levels into the A/D, with the subjects positioned as directed by the experimenter. Once set, the gains were not adjusted during the recording session.

The PC-based direct-to-disk recording system consisted of a Spectral Synthesis AudioPrisma board, which controlled recording and playback operations and stored the digitized audio data on a dedicated PC-based hard drive. The audio data were digitized and stored using the AudioPrisma board and a Spectral Synthesis ADAX-8818 Digital Audio Converter (A/D D/A). The ADAX-8818 is a Multi-channel (8-channel-in and 8-channel-out) digital audio converter.

A Pentium 90-MHZ. PC was used as the host for the AudioPrisma system, a 2 GB. magnetic disk drive, and the data collection software.

Data Collection Software

An in-house-developed Microsoft Windows-based program was used to prompt the subjects and collect their speech. The program uses a "push and hold-to-record" scheme for recording.
After recording an utterance, the program permits subjects to play back the waveform, and "Redo" or "Accept" the recording. When a subject accepts a recording, the next prompt is automatically displayed. Data collection continues until the prompt list is exhausted. At the beginning and ending of a data collection session, subjects are prompted to be quiet while the system samples the background noise. A short (nominally 5 second) recording of the ambient noise is then made using identical settings and parameters as for the speech recordings.

Data Preparation and Down Sampling

The data are transferred via NFS from the data drive on the host PC to a hard drive on a Sun Sparc20 workstation. The data are then down-sampled, from 32 kHz to 16 kHz, using the Entropic Waves+ "sf_convert" function. This function converts the sampling frequency by using a low-pass interpolation filter designed by the Kaiser windowing method with a cutoff at half the final sampling rate.

After down-sampling, NIST SPHERE headers are prepended to the waveform files, and the waveforms are debiased to remove DC offset. Finally, the waveforms are compressed using the SPHERE-embedded "Shorten" algorithm.

APPENDIX 2.

Secondary Microphones for the Test Set

1. A Shure SM58 Microphone was used for Test Subset 1 (channel b of the data).

This microphone is a professional quality cardioid dynamic microphone sometimes advertised as "the world standard professional stage microphone". The manufacturer's literature describes it as "a rugged unidirectional dynamic vocal microphone with a highly effective built-in wind and pop filter... a genuine world standard and a true audio legend". The Shure SM58 was mounted in an elastic shock mount, and suspended on a boom.

"Hum pickup" (typical)" is specified as 32 dB equivalent SPL in a 1 milliOersted field (60 Hz). A 1995-1996 catalog lists the price of this microphone as $188.75.

2. An Audio Technica AT851a Micro Cardioid Condenser Boundary Microphone was used for Test Subset 2 (channel f of the data).

This microphone is described in the manufacturer's literature as a "wide-range condenser microphone with a hemi-cardioid (half-space cardioid) polar pattern" with "a fixed-charge back plate permanently polarized condenser"... "useful in "surface-mounted applications such as high-quality sound reinforcement, professional recording, television, and other demanding applications". This microphone has switchable low-frequency roll-off settings. For this system, it was set to "flat response". Phantom power was supplied by the mixer. This microphone was positioned on the desk top just to the left of the keyboard of the PC-based workstation. No specifications for "hum pickup" are provided. A 1995-1996 catalog lists the price of this microphone as $285.

3. A Radio Shack #33-1060 Omni Electret Microphone was used for Test Subset 3 (channel g of the data).

This microphone is described as "...ideal for those special 'live action' recordings at meetings or interviews, parties, and much more!" and "has 20 to 30,000 Hz response, for use with most [portable] recorders". It was used with a "useful slip-on desk stand and a handy windscreen for noise reduction", and was powered by an AAA battery. This microphone was placed just below another desk-stand-mounted microphone, on the base of that microphone stand. No specifications for "hum pickup" are provided. A 1995 Radio Shack catalog lists the price as $23.99.

APPENDIX 3.

Texts for two speakers, with very low, and very high perplexity respectively

Detailed Orthographic Transcriptions (.dot) for Speaker 71c (Low Perplexity)

U\. S\. factory orders posted a larger than expected decline in July as stockpiles of unsold goods grew for the tenth consecutive month (71ec0201)
July\'s one point three percent drop in factory orders to a seasonally adjusted two hundred ninety two point nine one one billion dollars followed a revised point one percent loss in June (71ec0202)
Previously the government reported a point two percent decrease in June orders (71ec0203)
July\'s overall decline was the largest since April and the fifth decrease in six months the Commerce Department said (71ec0204)
While much of July\'s drop reflected the seasonal shutdown of auto plants for model year changeover excluding transportation orders still showed a point three percent loss for the month (71ec0205)
Industrial machinery and metals orders also declined (71ec0206)
Many businesses are still trying to liquidate inventories that built up after consumer spending stalled earlier in the year analysts said (71ec0207)
Factory inventories for example rose point six percent in July the government said (71ec0208)
Still some analysts believe manufacturing activity has stabilized after a bumpy first half of the year and the loss of more than a hundred and forty thousand factory jobs (71ec0209)
We\'re poised to do a little better Raymond Stone a managing director at Stone and McCarthy Research Associates said earlier this week in a Bloomberg Forum (71ec020a)
However I don\'t think it\'s going to boom Stone said (71ec020b)
Before today\'s Commerce Department report economists expected a point eight decline in factory orders during July according to a survey by Bloomberg Business News (71ec020c)
Excluding the defense industry factory orders decreased one point two percent in July (71ec020d)
Orders for durable goods which are made to last three or more years dropped two point one percent in July (71ec020e)
Preliminary July figures released a week ago showed a one point seven percent loss in orders for durable goods (71ec020f)

Detailed Orthographic Transcriptions (.dot) for Speaker 71f (High Perplexity)

H:ow does your boss hate you (71fc0201)
Let me count the ways (71fc0202)
For Carol it was her resemblance to snooty waitress Diane Chambers of T\. V\.\'s Cheers a character her boss detested (71fc0203)
Vicki\'s boss took out on her the frustrations of an unrequited love (71fc0204)
Barbara\'s boss didn\'t like her assertive style (71fc0205)
Many of us have had bosses who didn\'t like us for reasons that {ranged/range} from silly to serious (71fc0206)
Career experts insist you shouldn\'t take it personally but how can you not (71fc0207)
What\'s more personal or career threatening than this (71fc0208)
The usual resolutions put up with it quit or get fired can dent your self-confidence (71fc0209)
But some soured relationships can be mended or at least made tolerable and some bosses *self-destruct* (71fc020a)
Even if you must leave you can do it in a way that keeps your career intact (71fc020b)
None of this is easy and advice from career experts can be contradictory (71fc020c)
Challenge the boss says one (71fc020d)
The boss is always right even when wrong insists another (71fc020e)
Some say you can go over your boss\'s head while others say that\'s career suicide (71fc020f)

Nov 95 Hub and Spoke CSR Evaluation
Hub 3: Multi-microphone

GOAL: Improve basic SI performance on clean and unknown secondary channel
DATA: 20 spkrs * 15 Utts = 300 utts (Sennheiser)

20 spkrs * 15 utts = 300 utts (Non-Sennheiser)

Primary and Contrast Conditions

P0 (req) Open Baseline, compensation enabled test: using utts from non-Sennheiser 410 Mic.

CO (req) Open Baseline, compensation enabled test: using utts from Sennheiser 410 Mic.

C1A (opt) Controlled Eval Contrast - Static SI Test: Using SI-284 or SI-37 training sets, 60K-word Tri-gram, non-Sennheiser 410 Mic.

C1B (opt) Same System as C1A, but on the Sennheiser Mic.

C2A (opt) Static SI Test with compensation disabled using utts from Non-Sennheiser 410 Mic.

C2B (opt) Static SI Test with compensation disabled using utts from Sennheiser Mic.

C3A (opt) Same test as H0, except text material up to 1994 was used for the LM and lexicon, Non-Sennheiser 410 Mic.

C3B (opt) Same test as C0, except text material up to 1994 was used for the LM and lexicon, Sennheiser 410 Mic.

C4A (opt) Same test as H0, except using supervised adaptation with previous utterance, Non-Sennheiser 410 Mic.

C4B (opt) Same test as C0, except using supervised adaptation with previous utterance, Sennheiser 410 Mic.

SIDE INFO: Article boundaries unknown for: c1a c1b c2a c2b

	Primary P0	Contrast C0	Contrast C1A	Contrast C1B	Contrast C2A	Contrast C2B	Contrast C3A	Contrast C3B	Contrast C4A	Contrast C4B
System	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)
att1	55.5	10.6								
att2	55.5 *	9.5 *								
cmu2	29.0	13.6	27.3	14.2	68.3	14.2	32.7	15.3		
cmu2b			29.4	14.2	68.3 #	14.2 @				
cu-con1	19.8	12.5	24.2	14.0	24.2 $	14.0 &				
cu-htk1	13.5	6.6								
limsi1	17.8	9.3								
limsi1b	17.5 *	9.1 *								
limsi2	17.5 +	8.6 +								
nyu-sri1	24.0	9.4							24.3	9.3
ru1	55.0 +	20.2	55.0 =	20.2 ~						
sri1	24.6	9.7								

* Late/Debugged
$ Same as cu-con1 C1A
= Same as ru1 P0
Same as cmu2 C2A
& Same as cu-con1 C1B
~ Same as ru1 C0
@ Same as cmu2 C2B
+ Late

Note: In these tests, at both CMU and CU-CON, the C1 and C2 systems did not use compensation for the Sennheiser mic. data. Thus the data shown for Contrast C2B is identical to that shown for Contrast C1B.

COMPARISONS AND SIGNIFICANCE TESTS

	Test	% Increase	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
att1	P0:CO	424.1%	CO	CO	CO	CO
att2	P0:CO	486.6%	CO	CO	CO	CO
cmu2	P0:CO	113.3%	CO	CO	CO	CO
cu-con1	P0:CO	58.7%	CO	CO	CO	CO
cu-htk1	P0:CO	103.5%	CO	CO	CO	CO
limsi1	P0:CO	91.1%	CO	CO	CO	CO
limsi1b	P0:CO	91.5%	CO	CO	CO	CO
limsi2	P0:CO	102.9%	CO	CO	CO	CO
nyu-sri1	P0:CO	155.9%	CO	CO	CO	CO
ru1	P0:CO	171.9%	CO	CO	CO	CO
sri1	P0:CO	153.4%	CO	CO	CO	CO

	Test	% Increase	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	C1A:C1B	92.5%	C1B	C1B	C1B	C1B
cmu2b	C1A:C1B	107.9%	C1B	C1B	C1B	C1B
cu-con1	C1A:C1B	72.0%	C1B	C1B	C1B	C1B
ru1	C1A:C1B	171.9%	C1B	C1B	C1B	C1B

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	P0:C1A	-6.5%	same	C1A	same	same
cu-con1	P0:C1A	18.1%	P0	P0	P0	P0
ru1	P0:C1A	0.0%	same	same	same	same

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	CO:C1B	3.9%	same	same	same	same
cu-con1	CO:C1B	11.3%	same	CO	CO	CO
ru1	CO:C1B	0.0%	same	same	same	same

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	C1A:C2A	60.1%	C1A	C1A	C1A	C1A
cmu2b	C1A:C2A	56.9%	C1A	C1A	C1A	C1A
cu-con1	C1A:C2A	0.0%	same	same	same	same

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	C1B:C2B	0.0%	same	same	same	same
cmu2b	C1B:C2B	0.0%	same	same	same	same
cu-con1	C1B:C2B	0.0%	same	same	same	same

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	P0:C3A	11.1%	P0	P0	P0	P0

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2	CO:C3B	11.1%	same	CO	same	CO

Table 1.

Composite Report of All Significance Tests
For the Nov 95 ARPA CSR Hub 3 - C0 Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error)	MP
Signed Paired Comparison (Speaker Word Error Rate (%))	SI
Wilcoxon Signed Rank (Speaker Word Error Rate (%))	WI
McNemar (Sentence Error)	MN

	att1	att2	cmu2	cu-con1	cu-htk1	limsi1	limsi1b	limsi2	nyu-sri1	ru1	sri1
att1		MP att2 SI att2 WI att2 MN same	MP att1 SI same WI att1 MN att1	MP att1 SI same WI same MN att1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI same WI limsi1 MN limsi1	MP limsi1b SI same WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI same WI nyu-sri1 MN same	MP att1 SI att1 WI att1 MN att1	MP same SI same WI same MN same
att2			MP att2 SI att2 WI att2 MN att2	MP att2 SI att2 WI att2 MN att2	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP att2 SI att2 WI att2 MN att2	MP same SI same WI same MN same
cmu2				MP cu-con1 SI same WI same MN same	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP cmu2 SI cmu2 WI cmu2 MN cmu2	MP sri1 SI sri1 WI sri1 MN sri1
cu-con1					MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI same WI limsi1 MN limsi1	MP limsi1b SI same WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP cu-con1 SI cu-con1 WI cu-con1 MN cu-con1	MP sri1 SI sri1 WI sri1 MN sri1
cu-htk1						MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN same	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1
limsi1							MP same SI same WI same MN same	MP limsi2 SI limsi2 WI limsi2 MN same	MP same SI same WI same MN same	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP same SI same WI same MN same
limsi1b								MP limsi2 SI same WI limsi2 MN same	MP same SI same WI same MN same	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP same SI same WI same MN same
limsi2									MP same SI same WI same MN same	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP limsi2 SI same WI limsi2 MN limsi2
nyu-sri1										MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP same SI same WI same MN same
ru1											MP sri1 SI sri1 WI sri1 MN sri1
sri1											

Table 2.

Composite Report of All Significance Tests
For the Nov 95 ARPA CSR Hub 3 - P0 Test

Test Name	Abbrev.
matched pair sentence segment (word error)	mp
signed paired comparison (speaker word error rate (%))	si
wilcoxon signed rank (speaker word error rate (%))	wi
mcnemar (sentence error)	mn

	att1	att2	cmu2	cu-con1	cu-htk1	limsi1	limsi1b	limsi2	nyu-sri1	ru1	sri1
att1		MP same SI same WI same MN same	MP cmu2 SI cmu2 WI cmu2 MN cmu2	MP cu-con1 SI cu-con1 WI cu-con1 MN cu-con1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP same SI same WI same MN att1	MP sri1 SI sri1 WI sri1 MN sri1
att2			MP cmu2 SI cmu2 WI cmu2 MN cmu2	MP cu-con1 SI cu-con1 WI cu-con1 MN cu-con1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP same SI same WI same MN att2	MP sri1 SI sri1 WI sri1 MN sri1
cmu2				MP cu-con1 SI cu-con1 WI cu-con1 MN same	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP nyu-sri1 SI same WI nyu-sri1 MN nyu-sri1	MP cmu2 SI cmu2 WI cmu2 MN cmu2	MP sri1 SI same WI sri1 MN sri1
cu-con1					MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP limsi1 SI same WI same MN limsi1	MP limsi1b SI same WI same MN limsi1b	MP limsi2 SI same WI same MN limsi2	MP cu-con1 SI cu-con1 WI cu-con1 MN same	MP cu-con1 SI cu-con1 WI cu-con1 MN cu-con1	MP cu-con1 SI cu-con1 WI cu-con1 MN same
cu-htk1						MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1	MP cu-htk1 SI cu-htk1 WI cu-htk1 MN cu-htk1
limsi1							MP limsi1b SI same WI limsi1b MN same	MP limsi2 SI same WI limsi2 MN same	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1 SI limsi1 WI limsi1 MN limsi1	MP limsi1 SI limsi1 WI limsi1 MN limsi1
limsi1b								MP same SI same WI same MN same	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi1b SI limsi1b WI limsi1b MN limsi1b	MP limsi1b SI limsi1b WI limsi1b MN limsi1b
limsi2									MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP limsi2 SI limsi2 WI limsi2 MN limsi2	MP limsi2 SI limsi2 WI limsi2 MN limsi2
nyu-sri1										MP nyu-sri1 SI nyu-sri1 WI nyu-sri1 MN nyu-sri1	MP nyu-sri1 SI same WI same MN same
ru1											MP sri1 SI sri1 WI sri1 MN sri1
sri1											

Table 3.

Nov 95 Hub and Spoke CSR Evaluation
Hub 3: Multi-microphone

GOAL: Improve basic SI performance on clean and unknown secondary channel
DATA: 20 spkrs * 15 Utts = 300 utts (Sennheiser)

20 spkrs * 15 utts = 300 utts (Non-Sennheiser)

Primary and Contrast Conditions

P0 (req) Open Baseline, compensation enabled test: using utts from non-Sennheiser 410 Mic.

C0 (req) Open Baseline, compensation enabled test: using utts from Sennheiser 410 Mic.

C1A (opt) Controlled Eval Contrast - Static SI Test: Using SI-284 or SI-37 training sets, 60K-word Tri-gram, non-Sennheiser 410 Mic.

C1B (opt) Same System as C1A, but on the Sennheiser Mic.

C2A (opt) Static SI Test with compensation disabled using utts from Non-Sennheiser 410 Mic.

C2B (opt) Static SI Test with compensation disabled using utts from Sennheiser Mic.

C3A (opt) Same test as H0, except text material up to 1994 was used for the LM and lexicon, Non-Sennheiser 410 Mic.

C3B (opt) Same test as C0, except text material up to 1994 was used for the LM and lexicon, Sennheiser 410 Mic.

C4A (opt) Same test as H0, except using supervised adaptation with previous utterance, Non-Sennheiser 410 Mic.

C4B (opt) Same test as C0, except using supervised adaptation with previous utterance, Sennheiser 410 Mic.

SIDE INFO: Article boundaries unknown for: c1a c1b c2a c2b

	Primary P0	Contrast C0	Contrast C1A	Contrast C1B	Contrast C2A	Contrast C2B	Contrast C3A	Contrast C3B	Contrast C4A	Contrast C4B
System	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)
att1_set1	36.0	9.6								
att1_set2	55.3	9.4								
att1_set3	81.2	13.2								
att2_set1	36.0 *	8.7 *								
att2_set2	55.3 *	8.1 *								
att2_set3	81.2 *	11.9 *								
cmu2_set1	18.8	11.2	19.3	10.6	49.2	10.6	21.3	13.2		
cmu2_set2	28.0	11.8	23.9	12.4	71.9	12.4	32.4	13.7		
cmu2_set3	43.4	18.6	41.2	20.7	89.5	20.7	47.7	19.7		
cmu2b_set1			17.3	10.6	49.2 #	10.6 @				
cmu2b_set2			27.6	12.4	71.9 #	12.4 @				
cmu2b_set3			47.2	20.7	89.5 #	20.7 @				
cu-con1_set1	10.7	9.2	15.1	11.1	15.1 $	11.1 &				
cu-con1_set2	18.3	11.8	22.6	12.5	22.6 $	12.5 &				
cu-con1_set3	33.1	17.4	37.6	19.5	37.6 $	19.5 &				
cu-htk1_set1	8.1	5.1								
cu-htk1_set2	10.3	5.4								
cu-htk1_set3	23.9	9.9								
limsi1_set1	11.7	7.3								
limsi1_set2	14.3	7.8								
limsi1_set3	29.5	13.6								
limsi1b_set1	11.6 *	7.4 *								
limsi1b_set2	14.2 *	7.6 *								
limsi1b_set3	28.7 *	13.1 *								
limsi2_set1	11.6 +	6.6 +								
limsi2_set2	14.2 +	6.8 +								
limsi2_set3	28.7 +	13.2 +								
nyu-sri1_set1	13.2	6.7							13.5	6.7
nyu-sri1_set2	21.4	8.7							22.0	8.4
nyu-sri1_set3	40.7	13.6							40.8	13.7
ru1_set1	44.3 +	16.7	44.3 =	16.7 -						
ru1_set2	52.4 +	19.3	52.4 =	19.3 ~						
ru1_set3	71.6 +	25.9	71.6 =	25.9 ~						
sri1_set1	13.4	7.1								
sri1_set2	22.3	8.8								
sri1_set3	41.7	14.1								

* Late/Debugged
$ Same as cu-con1 C1A
= Same as ru1 P0

# Same as cmu2 C2A
& Same as cu-con1 C1B
~ Same as ru1 C0

@ Same as cmu2 C2B
+ Late

Note: In these tests, at both CMU and CU-CON, the C1 and C2 systems did not use compensation for the Sennheiser mic. data. Thus the data shown for Contrast C2B is identical to that shown for Contrast C1B.

Table 4 (part a).

COMPARISONS AND SIGNIFICANCE TESTS

	Test	% Increase		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
att1_set1	P0:CO	275.3%	CO	CO	CO	CO
att1_set2	P0:CO	486.9%	CO	CO	CO	CO
att1_set3	P0:CO	516.9%	CO	CO	same	CO
att2_set1	P0:CO	312.3%	CO	CO	CO	CO
att2_set2	P0:CO	581.6%	CO	CO	CO	CO
att2_set3	P0:CO	583.9%	CO	CO	same	CO
cmu2_set1	P0:CO	67.8%	CO	CO	CO	CO
cmu2_set2	P0:CO	136.7%	CO	CO	CO	CO
cmu2_set3	P0:CO	133.1%	CO	CO	same	CO
cu-con1_set1	P0:CO	17.4%	same	CO	same	same
cu-con1_set2	P0:CO	55.1%	CO	CO	same	CO
cu-con1_set3	P0:CO	89.6%	CO	CO	same	CO
cu-htk1_set1	P0:CO	58.8%	CO	CO	CO	CO
cu-htk1_set2	P0:CO	88.5%	CO	CO	CO	CO
cu-htk1_set3	P0:CO	142.4%	CO	CO	same	CO
limsi1_set1	P0:CO	59.6%	CO	CO	same	CO
limsi1_set2	P0:CO	83.9%	CO	CO	CO	CO
limsi1_set3	P0:CO	117.7%	CO	CO	same	CO
limsi1b_set1	P0:CO	57.0%	CO	CO	same	CO
limsi1b_set2	P0:CO	86.9%	CO	CO	CO	CO
limsi1b_set3	P0:CO	119.7%	CO	CO	same	CO
limsi2_set1	P0:CO	75.3%	CO	CO	CO	CO
limsi2_set2	P0:CO	110.1%	CO	CO	CO	CO
limsi2_set3	P0:CO	116.9%	CO	CO	same	CO
nyu-sri1_set1	P0:CO	96.8%	CO	CO	CO	CO
nyu-sri1_set2	P0:CO	146.4%	CO	CO	CO	CO
nyu-sri1_set3	P0:CO	200.2%	CO	CO	same	CO
ru1_set1	P0:CO	166.0%	CO	CO	CO	CO
ru1_set2	P0:CO	171.9%	CO	CO	CO	CO
ru1_set3	P0:CO	176.9%	CO	CO	same	CO
sri1_set1	P0:CO	89.1%	CO	CO	CO	CO
sri1_set2	P0:CO	153.0%	CO	CO	CO	CO
sri1_set3	P0:CO	195.5%	CO	CO	same	CO
	Test	% Increase		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	C1A:C1B	82.1%	C1B	C1B	C1B	C1B
cmu2_set2	C1A:C1B	93.6%	C1B	C1B	C1B	C1B
cmu2_set3	C1A:C1B	98.8%	C1B	C1B	same	C1B
cmu2b_set1	C1A:C1B	63.8%	C1B	C1B	C1B	C1B
cmu2b_set2	C1A:C1B	122.9%	C1B	C1B	C1B	C1B
cmu2b_set3	C1A:C1B	127.5%	C1B	C1B	same	C1B
cu-con1_set1	C1A:C1B	35.9%	C1B	C1B	C1B	C1B
cu-con1_set2	C1A:C1B	80.3%	C1B	C1B	C1B	C1B
cu-con1_set3	C1A:C1B	93.1%	C1B	C1B	same	C1B
ru1_set1	C1A:C1B	166.0%	C1B	C1B	C1B	C1B
ru1_set2	C1A:C1B	171.9%	C1B	C1B	C1B	C1B
ru1_set3	C1A:C1B	176.9%	C1B	C1B	same	C1B
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	P0:C1A	2.2%	same	same	same	same
cmu2_set2	P0:C1A	-17.1%	same	C1A	same	same
cmu2_set3	P0:C1A	-5.3%	same	same	same	same
cu-con1_set1	P0:C1A	29.0%	P0	P0	P0	P0
cu-con1_set2	P0:C1A	19.0%	same	P0	P0	P0
cu-con1_set3	P0:C1A	11.9%	same	P0	same	P0
ru1_set1	P0:C1A	0.0%	same	same	same	same
ru1_set2	P0:C1A	0.0%	same	same	same	same
ru1_set3	P0:C1A	0.0%	same	same	same	same
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	CO:C1B	-6.1%	same	same	same	same
cmu2_set2	CO:C1B	4.2%	same	same	same	same
cmu2_set3	CO:C1B	10.2%	same	CO	same	CO
cu-con1_set1	CO:C1B	17.8%	same	CO	same	CO
cu-con1_set2	CO:C1B	5.9%	same	same	same	same
cu-con1_set3	CO:C1B	10.3%	same	CO	same	CO
ru1_set1	CO:C1B	0.0%	same	same	same	same
ru1_set2	CO:C1B	0.0%	same	same	same	same
ru1_set3	CO:C1B	0.0%	same	same	same	same
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	C1A:C2A	60.9%	C1A	C1A	C1A	C1A
cmu2_set2	C1A:C2A	66.7%	C1A	C1A	C1A	C1A
cmu2_set3	C1A:C2A	53.9%	C1A	C1A	same	C1A
cmu2b_set1	C1A:C2A	64.8%	C1A	C1A	C1A	C1A
cmu2b_set2	C1A:C2A	61.7%	C1A	C1A	C1A	C1A
cmu2b_set3	C1A:C2A	47.3%	same	C1A	same	C1A
cu-con1_set1	C1A:C2A	0.0%	same	same	same	same
cu-con1_set2	C1A:C2A	0.0%	same	same	same	same
cu-con1_set3	C1A:C2A	0.0%	same	same	same	same
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	C1B:C2B	0.0%	same	same	same	same
cmu2_set2	C1B:C2B	0.0%	same	same	same	same
cmu2_set3	C1B:C2B	0.0%	same	same	same	same
cmu2b_set1	C1B:C2B	0.0%	same	same	same	same
cmu2b_set2	C1B:C2B	0.0%	same	same	same	same
cmu2b_set3	C1B:C2B	0.0%	same	same	same	same
cu-con1_set1	C1B:C2B	0.0%	same	same	same	same
cu-con1_set2	C1B:C2B	0.0%	same	same	same	same
cu-con1_set3	C1B:C2B	0.0%	same	same	same	same
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	P0:C3A	11.5%	same	P0	same	P0
cmu2_set2	P0:C3A	13.4%	same	P0	same	same
cmu2_set3	P0:C3A	9.1%	same	P0	same	P0
	Test	% Reduct.		Significance Tests:		
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
cmu2_set1	CO:C3B	15.3%	CO	CO	same	CO
cmu2_set2	CO:C3B	13.7%	same	CO	same	same
cmu2_set3	CO:C3B	5.4%	same	same	same	same

Table 4 (part b).

Subset 1

Spkr	Sex	Avg. Perplexity	% OOV Words	Speech Rate (words/min)	Number of Words	S/N "A" (dB)	S/N "B" (dB)	A-Weighted S/N "A"	A-Weighted S/N "B"
710	f	286	0.9	192	324	40.2	23.0	41.5	25.7
716	m	66	0.0	176	289	35.7	18.3	35.4	20.6
717	f	289	3.3	181	299	32.0	18.0	31.9	18.5
71c	f	10	0.8	172	400	36.2	15.0	36.2	16.1
71g	m	151	0.8	240	395	35.0	20.0	34.6	21.1
71h	m	165	0.0	188	317	41.1	20.7	41.6	22.4
71j	f	287	2.1	170	292	35.8	17.3	37.0	20.8
Means		179	1.129	188		36.6	18.9	36.9	20.7

Subset 2

Spkr	Sex	Avg. Perplexity	% OOV Words	Speech Rate (words/min)	Number of Words	S/N "A" (dB)	S/N "F" (dB)	A-Weighted S/N "A"	A-Weighted S/N "F"
712	m	73	0.0	217	330	34.2	17.3	34.0	16.9
714	m	539	0.6	158	310	42.8	26.4	43.2	27.1
719	f	120	0.7	162	277	37.6	16.1	35.2	15.9
71d	m	308	0.4	210	241	42.5	22.1	42.2	23.3
71e	f	84	0.0	197	297	36.7	21.1	35.6	21.9
71f	m	964	1.7	165	180	37.8	20.8	36.9	22.2
71i	f	418	0.7	174	267	38.4	19.5	38.7	21.1
Means		358	0.586	183		38.6	20.5	38.0	21.2

Subset 3

Spkr	Sex	Avg. Perplexity	% OOV Words	Speech Rate (words/min)	Number of Words	S/N "A" (dB)	S/N "G" (dB)	A-Weighted S/N "A"	A-Weighted S/N "G"
711	m	414	1.6	184	307	38.6	7.0	38.1	16.4
713	m	353	4.1	205	270	37.0	8.9	37.2	17.1
715	f	911	2.1	212	282	41.8	13.3	42.9	21.5
718	f	560	1.0	227	292	41.7	9.3	43.5	19.5
71a	f	135	0.0	187	314	32.2	10.5	32.5	19.2
71b	m	416	3.5	232	315	39.4	11.3	37.7	15.7
Means		465	2.05	208		38.5	10.1	38.7	18.2

Table 5.

~300 words x 20 Speakers ≐ 6000 words.

96 "Word Errors"

Say 100/6000 ~ 1/60 ≐ 1.6%

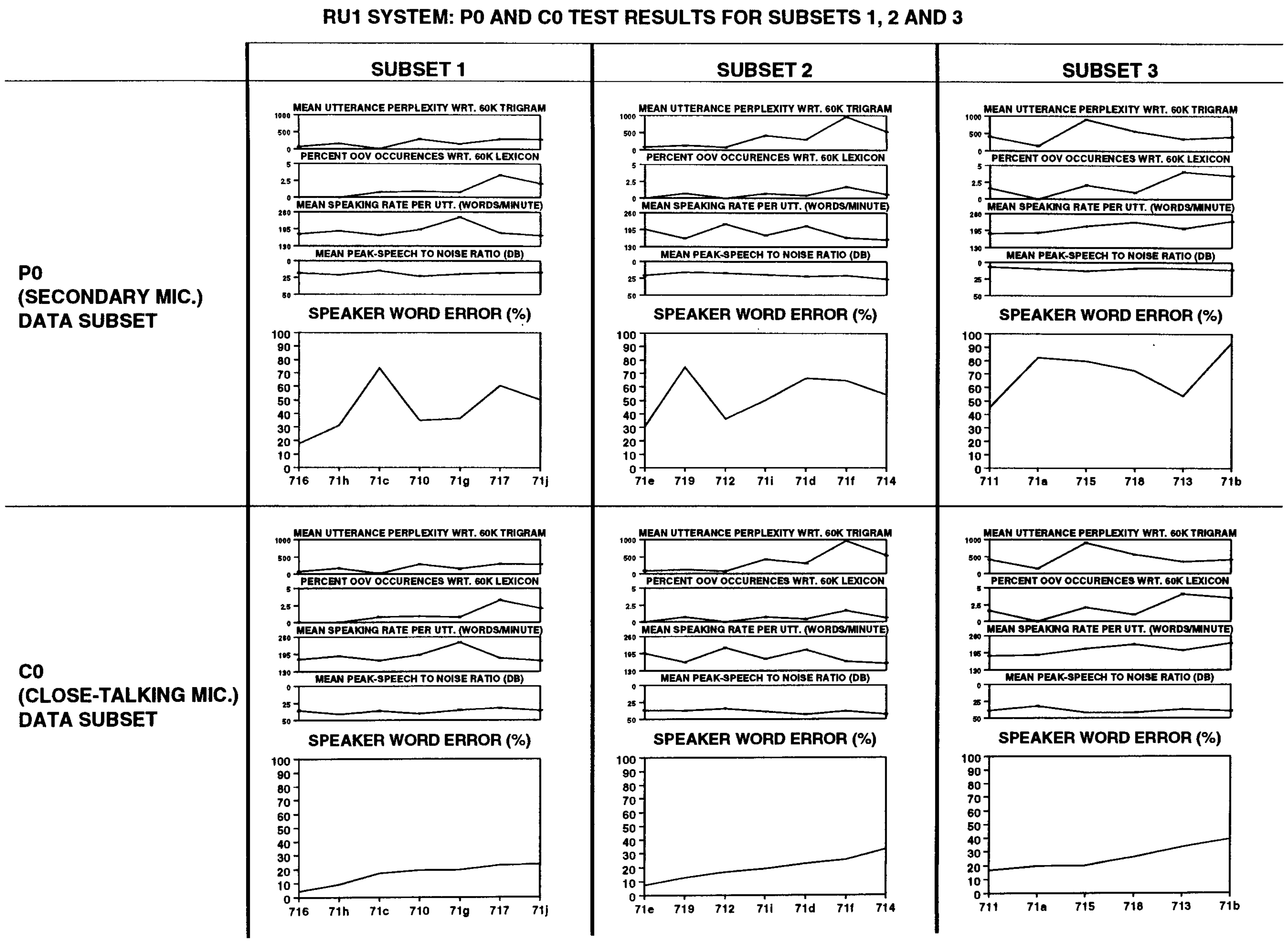
RU1 SYSTEM: P0 AND C0 TEST RESULTS FOR SUBSETS 1, 2 AND 3
SUBSET 1
SUBSET 2
SUBSET 3
P0
(SECONDARY MIC.)
DATA SUBSET
C0
(CLOSE-TALKING MIC.)
DATA SUBSET
MEAN UTTERANCE PERPLEXITY WRT. 60K TRIGRAM
1000
500
0
PERCENT OOV OCCURENCES WRT. 60K LEXICON
5
2.5
0
MEAN SPEAKING RATE PER UTT. (WORDS/MINUTE)
260
195
130
MEAN PEAK-SPEECH TO NOISE RATIO (DB)
0
25
50
SPEAKER WORD ERROR (%)
100
90
80
70
60
50
40
30
20
10
0
716
71h
71c
710
71g
717
71j
71e
719
712
71i
71d
71f
714
711
71a
715
718
713
71b

Figure 1.

CU-HTK1 SYSTEM: P0 AND C0 TEST RESULTS FOR SUBSETS 1, 2 AND 3

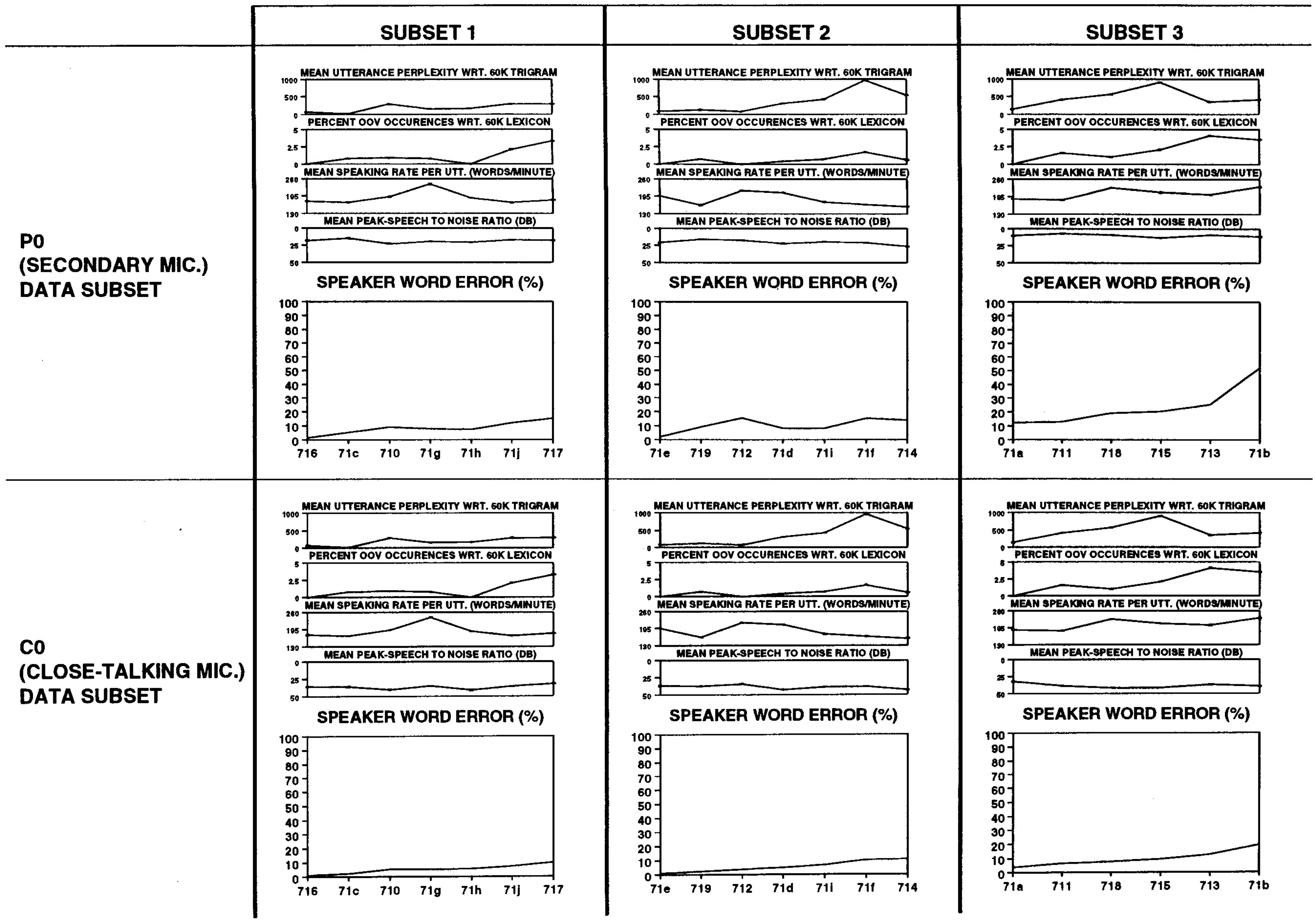

Figure 2.